Award Number: W81XWH-12-1-0323

TITLE: Advanced Lung Cancer Screening: An Individualized Molecular Nanotechnology Approach

PRINCIPAL INVESTIGATOR: James Herman, Professor

CONTRACTING ORGANIZATION: Johns Hopkins University, The
Baltimore, MD 21218-2680

REPORT DATE: March 2016

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE		Form Approved OMB No. 0704-0188
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**		
1. REPORT DATE March 2016	2. REPORT TYPE Final	3. DATES COVERED 1Aug2012 - 31Dec2015
4. TITLE AND SUBTITLE Advanced Lung Cancer Screening: An Individualized Molecular Nanotechnology Approach		5a. CONTRACT NUMBER W81XWH-12-1-0323 5b. GRANT NUMBER 5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) James G. Herman, MD email: hermanj3@upmc.edu		5d. PROJECT NUMBER 5e. TASK NUMBER 5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Johns Hopkins University, The Baltimore, MD 21218-2680		8. PERFORMING ORGANIZATION REPORT
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) **U.S. ARMY MEDICAL RESEARCH AND MATERIEL COMMAND FORT DETRICK. MARYLAND 21702-5012**		10. SPONSOR/MONITOR'S ACRONYM(S) 11. SPONSOR/MONITOR'S REPORT NUMBER(S)
12. DISTRIBUTION / AVAILABILITY STATEMENT Approved for Public Release; Distribution Unlimited		
13. SUPPLEMENTARY NOTES		

14. ABSTRACT

This grant utilizes complimentary approaches to improve the early detection of lung cancer. Our goal is to explore whether detection of DNA methylation changes and enhanced CT evaluations will add to the specificity of lung cancer detection. Based on our previous development of an improved panel of genes hypermethylated in lung cancer, with extraordinarily high specificity and sensitivity, we combined the improved methods of MOB with highly sensitive methylation specific PCR assays suitable for biologic fluid testing (sputum and serum) and completed the study of a cohort of cancer positive and negative samples. In combination with these molecular detection approaches, we have examined the alterations in air space for improving detection of lung cancer and find that variability of air spaces is associated with the presence of lung cancer. We have during the period of this grant developed a highly sensitive and specific method for early detection of lung cancer.

15. SUBJECT TERMS Nothing listed					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION	18. NUMBER	19a. NAME OF RESPONSIBLE PERSON USAMRMC
a. REPORT **Unclassified**	b. ABSTRACT Unclassified	c. THIS PAGE Unclassified	UU	40	19b. TELEPHONE NUMBER *(include area code)*

Standard Form 298
(Rev. 8-98)

Table of Contents

Page

1. INTRODUCTION:

This grant utilizes complimentary approaches to improve the early detection of lung cancer. Our goal is to explore whether detection of DNA methylation changes and enhanced CT evaluations will add to the specificity of lung cancer detection. This was defined in our aims.
Specific Aim 1: To improve the clinical utility and effectiveness of a nested, gel based DNA methylation assay for sputum and plasma by increasing its sensitivity and specificity through nanotechnology.
Hypothesis: Detection of DNA methylation from individuals with cancer can be used to determine lung cancer risk and can be enhanced through discovery of optimal hypermethylated genes and implementation of enhanced detection technologies.
Specific Aim 2: To use an in vitro molecular testing of sputum and serum with DNA methylation rather than simple demographics alone to select the highest risk smokers for an expensive screening modality such as CT scanning. Hypothesis: DNA methylation testing is more specific in selecting those at the highest risk for lung cancer than clinical demographics alone.
Specific Aim 3: To optimize low dose chest CT screening for lung cancer. Hypothesis: Valuable information on the chest CT scan, based on the severity, distribution, and pattern of low attenuation areas ("emphysema"), may be crucial to increasing our insights and effectiveness of determining lung cancer risk, the frequency of follow up scans, reducing false positives, and controlling costs compared to an annual chest CT screening for the sole use to detect lung cancer tumors after they occur.

2. KEYWORDS:

Lung Cancer Screening, CT Screening, DNA Methylation Detection, Emphysema Score, Lung Airspace Variability Score.

3. OVERALL PROJECT SUMMARY:

During the first two years of this proposal, we had largely accomplished the goals of building an improved method for methylation detection which comprise specific aim 1. We made significant progress on the two sub-aims of this proposal in implementing the developments from last year. Last year's progress included A) Developing optimal hypermethylated gene panels for detection of tumor DNA from lung cancer and B) Optimize nanotechnology based detection of DNA methylation for increased sensitivity and specificity. The first efforts were initially focused on the development of an optimal gene panel for detection of lung cancer. After completion of these studies, we published the results earlier (1) with a summary provided here. Hypermethylation of CpG islands is a common and important alteration in the transition from normal to transformed cells. Following previously validated methods for the discovery of cancer-specific hypermethylation changes from NSCLC cell lines, we identified >300 candidate genes. Using the Cancer Genome Atlas (TCGA) and employing extensive filtering to refine our candidate genes for the greatest ability to distinguish tumor from normal, we had initially defined a three-gene panel, CDO1, HOXA9, and TAC1, which we subsequently validate in two independent cohorts of primary NSCLC samples. This 3-gene panel is 100% specific, showing no methylation in 75 TCGA and 7 primary samples and is 83-99% sensitive for NSCLC (shown in last year's progress report). This panel has been further expanded through the identification of additional genes with extremely high methylation frequencies in lung cancer. This panel now includes three additional genes, HOXA7, SOX17 and ZFP42, for which real-time MSP analyses assays were also developed to complement the previous 3 gene panel to provide redundant tumor coverage to optimize detection. Our subsequent development of this panel expanded into 6 genes with good sensitivity and specificity, with results shown for 5 genes.

Figure 1. Methylation of *CDO1, HOXA9, SOX17, ZPF42* and *TAC1* is Highly Sensitive for NSCLC Detection, in stage I lung cancer samples. Highly prevalent methylation sites were chosen from data generated within the TCGA studies. All normal lung tissue lack DNA methylation, but the majority of lung tumors have methylation of individual loci, and overall, nearly all tumors have methylation of at least one loci (adapted from Wrangle (1) .

These new assays were confirmed to specifically detect abnormal methylation using normal lymphocytes and *in vitro* methylated bisulfite converted DNA. We found high specificity to methylation in bisulfite converted DNA and no amplification in unconverted and no template controls. These new assays were deployed in specific aim 2 with good results.

Aim 2: The use of methylated tumor-specific circulating DNA has shown great promise as a potential cancer biomarker. Nonetheless, the relative scarcity of tumor-specific circulating DNA presents a challenge for traditional DNA extraction and processing techniques. We completed a study of improvements in DNA processing, with a single tube extraction and processing technique dubbed "methylation on beads" that allows for DNA extraction and bisulfite conversion for up to 2 ml of plasma or serum (Outline of approach shown in figure 2) (2). In comparison to traditional techniques such as phenol chloroform and alcohol extraction, methylation on beads yields a 1.5 to 5-fold improvement in extraction efficiency. The greatest enhancement in extraction efficiency is seen with small amounts of DNA, precisely matching the need for improved extraction in low DNA content samples such as plasma and serum. A summary of the final results using this approach is provided in figure 3.

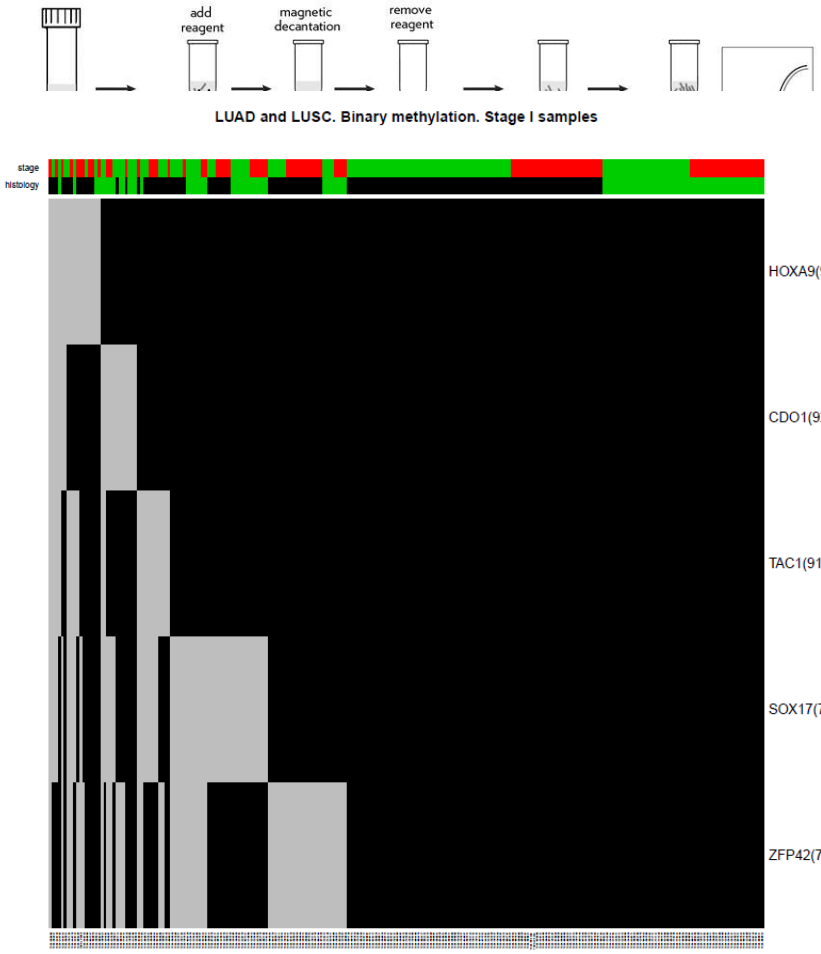


Figure 2. Overview of the Methylation-on-Beads (MOB) Process. Circulating DNA from up to 2 ml of plasma is extracted and purified via SSBs. The purified DNA is then subject to bisulfite conversion and analyzed via methylation specific PCR (MSP). The entire sample preparation process can be performed in a single tube and consists of an iterative process of adding reagents, magnetic decantation, and removal of supernatant.

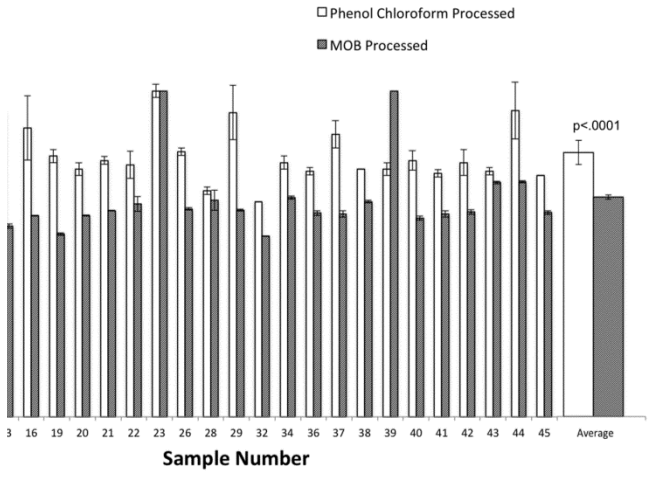


Figure 3 ß-Actin Ct values for MOB processed vs. Phenol Chloroform extracted and traditionally processed plasma samples from 24 patients diagnosed with lung cancer. The MOB technique demonstrates consistently higher and less variable recovery, as demonstrated by the lower average Ct value (33.8 vs. 40.6 cycles) and Ct standard deviation (0.3 vs. 1.9 cycles), respectively. This improvement in Ct of 6.8 cycles represents a $2^{6.8}$ or 111 fold increase in amplifiable DNA, on average.

Having developed an optimal panel and improved upon methods for processing the DNA as planned, we have applied these techniques to the plasma and serum of patients with CT detected lung cancer and those with non-cancerous nodules, and have completed the writing of a manuscript containing these results. The abstract of that manuscript is as follows (full manuscript in appendix):

Purpose:
To improve the diagnostic accuracy of lung cancer screening using ultrasensitive methods detecting gene promoter methylation in sputum and plasma using Methylation-On-Beads (MOB) with a lung cancer specific gene panel.
Patients and Methods:
This is a case-control study of subjects with nodules suspicious for lung cancer on CT imaging in which plasma and sputum were obtained pre-operatively. Cases (n=150) had pathological confirmation of node negative (stage IA, IB and IIA) non-small cell lung cancer while controls (n=60) had non-cancer diagnoses. We detected promoter methylation using quantitative methylation specific real-time PCR with MOB for cancer-specific genes (CDO1, TAC1, HOXA7, HOXA9, SOX17 and ZFP42) identified from The Cancer Genome Atlas (TCGA).
Results:
DNA methylation was detected in plasma and sputum more frequently in people with cancer compared to controls ($p<0.001$) for 5 of 6 genes examined. Individual gene detection The sensitivity and specificity for lung cancer diagnosis using individual genes from sputum ranged from 63-93% and 42-92% respectively and from plasma from 33-91% and 52-94%. A three-gene combination including the best individual genes has sensitivity and specificity of 93% and 79% using sputum and 91% and 64% using plasma. Area under the Receiver Operating Curve for this panel was 0.89 95% CI (0.80-0.98) in sputum and 0.77 95% CI (0.68-0.86) in plasma. Independent, blinded random forest prediction models combining gene methylation with age, pack-year, COPD status and FVC values correctly predicted lung cancer in 91% of subjects using sputum samples and 85% of subjects using blood samples.
Conclusions:
High diagnostic accuracy for early stage lung cancer can be obtained using methylated promoter detection in sputum or plasma.

Table 1. Baseline Characteristics of the 210 Subjects.

Patient Characteristics	Cancer (N=150)	Control (N=60)	*p* Value
Age at surgery (years) (IQR)	68 (62-75)	63 (55-73)	0.007
Gender			
Male (%)	63 (42%)	33 (55%)	0.094
Female (%)	87 (58%)	27 (45%)	
Race			
White (%)	120 (80%)	51 (85%)	
Black (%)	19 (13%)	3 (5%)	0.087
Other (%)	11 (7%)	6 (10%)	
Stage			
IA-IB (%)	136 (91%)	NA	NA
IIA (%)	14 (9%)	NA	
Histology			
Adenocarcinoma (%)	121 (81%)	NA	
Squamous-cell (%)	26 (17%)	NA	NA
Adenosquamous (%)	3 (2%)	NA	
Smoking status			
Current (%)	27 (18%)	7 (12%)	
Former (%)	87 (58%)	34 (57%)	0.176
Never (%)	31 (21%)	19 (32%)	
Pack-year (IQR)	30 (10-50)	20 (0-35)	0.010
COPD (%)	41 (27%)	12 (20%)	0.370
FEV1 % Predicted (IQR)	84 (70-99)	85 (70-100)	0.861
FVC % Predicted (IQR)	92 (80-103)	87 (80-110)	0.682
FEV1/FVC % Ratio (IQR)	73 (68-78)	77 (70-79)	0.080
Nodule size (cm)	2 (1.5-3)	1.5 (1.1-3)	0.01
< 1cm	6 (4%)	13 (22%)	
1-2 cm	52 (35%)	19 (32%)	0.001
> 2 cm	92 (61%)	28 (47%)	
Nodule volume (cm^3)	4.19 (1.77-14-14)	1.6 (0.52-18.12)	0.001

Abbreviations: Chronic obstructive pulmonary disease: COPD, Forced Expiratory Volume in one second: FEV1, Forced vital capacity: FVC, Interquartile range: IQR. Nodule size % <1cm, 1-2, >2cm

Methylation was readily detected for these loci in the majority of cancer patients, but not in most control patients. Actual quantitation of the methylation was carried out with the following DNA Methylation analysis: The genomic sequence for the genes and 1000 bases upstream was obtained from the UCSC genomic browser website. The primers and hybridization probes for methylation analysis were designed based on this sequence by using Primer3 (v.0.4.0). The analysis was performed using quantitative real-time Methylation Specific PCR, and normalized to a control β-Actin assay. Each reaction was performed in a 25 µl PCR mixture consisting of 2 µl of bisulfite converted DNA, 300 nM R-sense primer, 300 nM F-anti-sense primer, 100nM probe, 100 nM of fluorescein reference dye (Life Technologies), 1.67mM dNTPs (VWRQuotation), and 1 ul of Platinum Taq® DNA Polymerase (invitrogen). Master mix contained 16.6mM $(NH_4)_2SO_4$, 67mM Tris pH 8.8, 6.7 mM $MgCl_2$ and 10mM β-mercaptoethanol in a nuclease-free DI water solution. Amplification reactions were performed using 96 well-plates (MicroAmp®) with all samples being analyzed in triplicate. Thermo cycling conditions were as follows: 95°C for 5 min, 50 cycles at 95°C for 15 seconds, and 65°C for 1min and 72°C for 1 min. An ABI StepOnePlus Real-Time PCR system was used (Applied Bio Systems).

With the extremely low levels of DNA methylation in plasma and sputum, replicates for some samples produced no detectable methylation as expected. To incorporate this information into the final quantification of methylation, we calculated the $2^{-\Delta CT}$ for each methylation detection replicate comparing it to the mean Ct for β-Actin (ACTB). For replicates which were not detected (ND), a CT of 100 was used, creating a near zero value for $2^{-\Delta CT}$. The mean $2^{-\Delta CT}$ value was calculated with the formula:

$$\mu\, 2^{-\Delta CT} = \frac{(2^{-\Delta CT replicate\ 1} + 2^{-\Delta CT replicate\ 2} + 2^{-\Delta CT replicate\ 3})}{3}$$

The results of the methylation analyses for these six genes in the 210 patients are shown, with tumor methylation, sputum and plasma results.

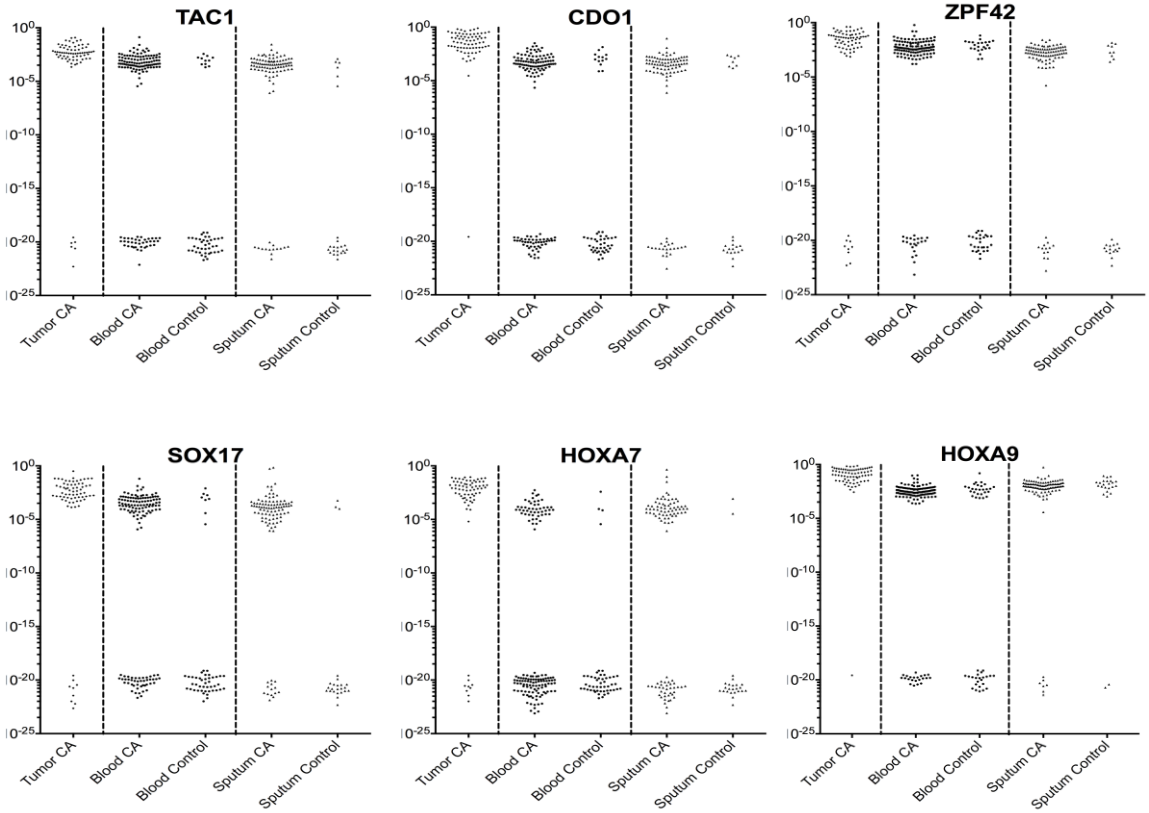


Figure 4. Methylation level (Normalized to beta actin) for 6 genes detected in tumor, plasma (blood) and sputum from patients with Lung cancer and non-cancer controls, plotted on log scale. Each dot represents the calculated level of methylation using the formula above from triplicates. Note the nearly universal detection of methylation in all tumor tissues, and at higher quantitative levels than seen in biologic fluids (as expected given the relative amounts of tumor DNA in tissue samples compared to plasma or sputum). Plasma and sputum samples vary in quantity of methylation from equal to that in tumor to very low level detection (10-5-10-6)_This low level detection was not possible without the integrated methods described for this proposal

With these results, we calculated the analytic accuracy of methylation detection.

Table 2. Gene Methylation Sensitivity, Specificity, AUC and Association with Cancer Diagnosis for genes obtained from Sputum and Blood.

Sputum	Sensitivity	Specificity	PPV	NPV	AUC	95% CI
CDO1	78%	67%	90%	45%	0.70	(0.57 - 0.84)
TAC1	84%	79%	94%	57%	0.84	(0.74 - 0.94)
HOXA7	63%	92%	97%	40%	0.77	(0.67 - 0.86)
HOXA9	77%	42%	83%	32%	0.56	(0.41 - 0.69)
SOX17	84%	88%	96%	59%	0.84	(0.75 -0.94)
ZFP42	88%	62%	90%	58%	0.73	(0.60 - 0.87)
TAC1, HOXA7, SOX17	93%	79%	94%	75%	0.89	(0.80 - 0.98)

Blood	Sensitivity	Specificity	PPV	NPV	AUC	95% CI
CDO1	65%	74%	86%	46%	0.68	(0.58 - 0.77)
TAC1	76%	78%	90%	57%	0.78	(0.70 - 0.86)
HOXA7	33%	94%	93%	36%	0.60	(0.51 - 0.69)
HOXA9	81%	52%	81%	52%	0.62	(0.52 - 0.73)
SOX17	71%	86%	93%	54%	0.78	(0.70 - 0.86)
ZFP42	81%	58%	83%	55%	0.66	(0.56 - 0.75)
CD01, TAC1, SOX17	91%	64%	86%	74%	0.77	(0.68 - 0.86)

Abbreviations: area under the curve (in the ROC curves): AUC, 95 % confidence interval: 95% CI.

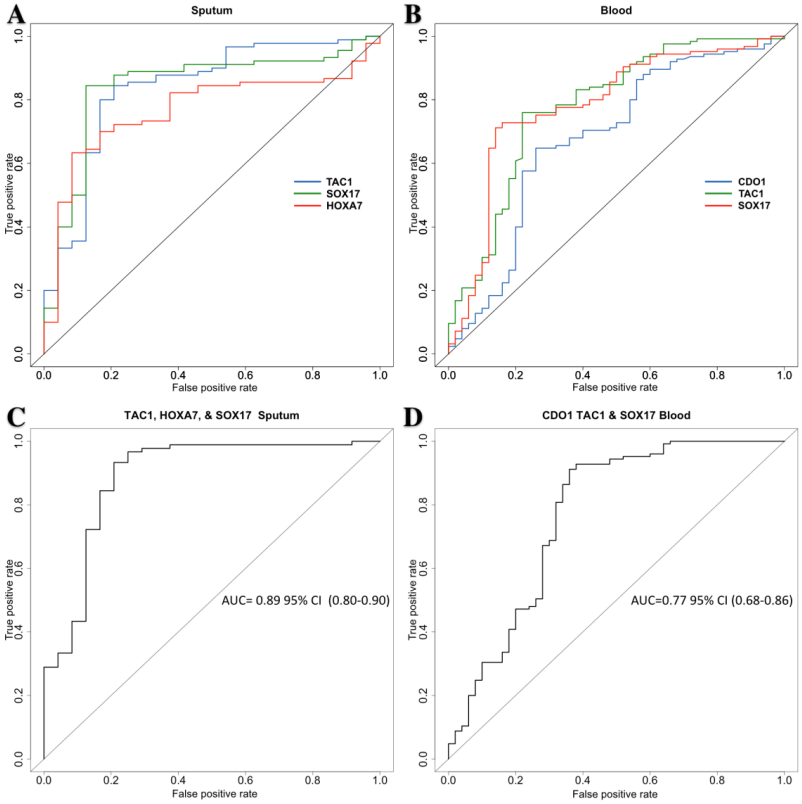


Figure 5. Receiver operator classification curves for lung cancer detection.
A. ROC curves comparing the 3 genes with the largest areas under the curve for sputum. **B.** ROC curves comparing the 3 genes with the largest areas under the curve for blood. **C.** ROC of the combined methylation status of the genes from sputum with the largest area under the curve. **D.** ROC of the combined methylation status of the genes from blood with the largest area under the curve.

Abbreviations: area under the curve: AUC, 95 % confidence interval: 95% CI.

Independent Prediction Accuracy Performance: While the above analysis looked at individual gene methylation in cases and controls to detect cancer, independent blinded random forest prediction models analyzed all these biomarkers in combination with clinical risk factors. Risk factors included in the first two random forest prediction models were methylation Ct values from all six genes, age, pack-year, COPD status and FVC values. The methylation Ct values were not included in the last prediction model. The randomly selected training dataset has 140 subjects with 99 (70.7%) cancers and 41 (29.3%) controls. The independent test set has 70 subjects with 51 (72.9%) cancers and 19 (27.1%) controls. In the variable of importance output of the first two random forest prediction models, methylation Ct values were ranked as more important variables than demographic and clinical variables (Figure 6). Table 3 summarizes the prediction accuracies of these three models when they were applied to the independent test set patients. With sputum samples, the random forest model correctly predicted lung cancer in 91% of subjects in the test subset. The corresponding AUC was 0.85 95% CI (0.59-1.0) . The sensitivity and specificity of the prediction in the testing subset from the ROC curve were 0.93 and 0.86, respectively. Using plasma samples, the random forest model correctly predicted lung cancer in 85% of subjects in the testing subset. The corresponding AUC was 0.89 95% CI (0.79-0.99). The sensitivity and specificity of the prediction in the testing subset from the ROC curve were 0.93 and 0.67, respectively. Using clinical and demographic risk factors alone, the accuracies were lower than the first two models with a diagnostic accuracy of 68%, AUC of 0.64, PPV of 75% and a NPV of 38% (Table 3).

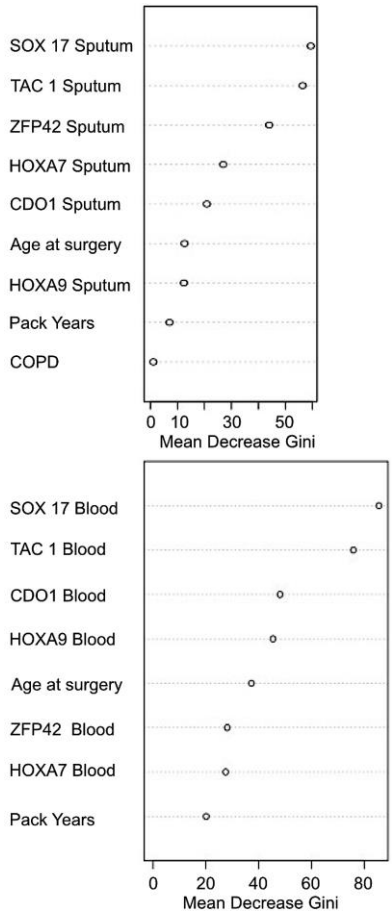


Figure 6. Variable importance plot for random forest prediction. The plot details the relative importance of each of the variables to the model's accuracy (including: methylation μ 2-ΔCT values, nodule size, age, pack-year, COPD status and FVC values). The x-axis is the mean decrease in the Gini co-efficient that results when that variable is included in the model. The Gini coefficient is a measure of inequality among the trees in the random forest, and in this case represents the performance of the random forest model with and without a variable included. Those variables that have the highest decrease in the Gini coefficient were most likely to create consensus among the individual decision trees used in the model (or reduce inequality) when included in the model. These variables are therefore most predictive of the outcome of the model overall. Those variables with a small decrease in the mean Gini coefficient are relatively less important to the prediction made by the random forest model.

Table 3. Performance for lung cancer diagnosis of the independent blinded random forest prediction models on the testing subset

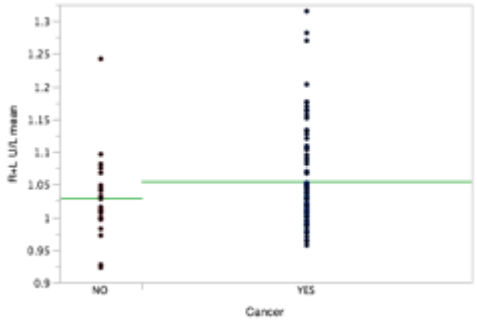
	Sensitivity	Specificity	PPV	NPV	AUC	95% CI
Prediction from Sputum	93%	86%	96%	75%	0.85	0.59-1
Prediction from Blood	93%	67%	87%	80%	0.89	0.79-0.99
Clinical Predictors alone	84%	26%	75%	38%	0.64	0.50-0.78

Abbreviations: area under the curve (in the ROC curves): AUC, 95 % confidence interval: 95% CI.

In the final period of our funding, the PI (Dr. Herman) moved to the University of Pittsburgh and has begun to implement this approach in samples from the Lung Cancer Pittsburgh Screening study (PLuSS) and the Pittsburgh Lung Cancer SPORE. This will form a validation cohort and be used to further improve this already promising approach.

For specific aim 3, to date we were able to identify 210 subjects in the SPORE database that had CT scans performed prior to surgery which were adequate for analysis. We have completed measurement of the extent of computed tomography (CT) in these subjects. Of the group, 168 of the subjects had cancer, and 42 did not.
The software can divide the lung into upper, middle, and lower fields on the right and the left for a total of six lung areas. For the subjects, clearly abnormal areas were eliminated from further analysis. For the Ca+ subjects, the final usable number of lung fields were right upper=106, right middle=111, right lower=108, left upper=118, left middle=118, and left lower=116. For the Ca- subjects, we have 103 for each lung field. The emphysema score was based on the number of voxels with Hounsfield units (HUs) less than -910. The percent emphysema of the lungs ranged from 0.19 to 56% among all the subjects with a mean score of 28.8±15% (mean±SD). The subjects with and without cancer had a similar amount of emphysema (29±15 and 27±14 respectively (p=0.42) . This suggests that simple screening for emphysema would not allow for detection of lung cancer.
We continued the study of CT images, examining lung heterogeneity by comparing the ratio of the upper to the lower lung in regards to the mean, standard deviation, <-950 HU, and 15^{th} percentile of the CT scans comparing cancer and non-cancer. The software divides the scans into right and left, allowing the examination of these 4 variables for each side separately. The only variable that was significant, and was significant on both the right (p=0.0091) and left (p=0.0454) was the 15^{th} percentile. The mean HU value for the right side was close (p=0.0785).
We next combined the right and left sides to have a single overall measure. (Figure 7) The average (right-left) upper to lower ratios for the mean CT score, standard deviation of the mean, <-950 HU, and 15^{th} percentile of the CT scans comparing the cancer to the non-cancer subjects. Again the average (right-left) upper to lower ratio for the 15^{th} percentile was significant (p=0.0014). In addition, the average (right-left) upper to lower ratio for the mean CT density was also significant (p=0.04) (Figure 8). Neither the standard deviation (p=0.43) or the <-950HU (p=0.26) measurements were significant.

Figure 7 Ratio (U/L) Mean CT density

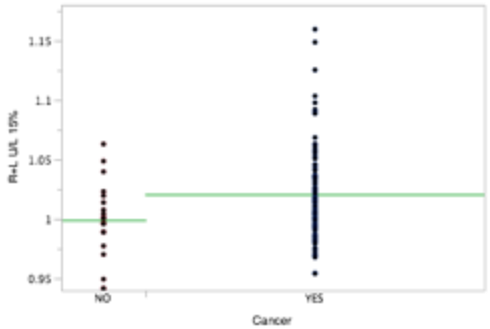


t Test; Assuming equal variances

Difference 0.025567 t Ratio 2.066335
Std Err Dif 0.012373 DF 208
Upper CL Dif 0.049961 Prob > |t| 0.0400*
Lower CL Dif 0.001174 Prob > t 0.0200*
Confidence 0.95 Prob < t 0.9800

These findings suggest that the upper to lower ratio of the mean intensity may be an independent predictor of lung cancer.

Figure 8. Ratio (U/L) 15th percentile



t Test; Assuming equal variances

Difference 0.021611 t Ratio 3.228527
Std Err Dif 0.006694 DF 208
Upper CL Dif 0.034807 Prob > |t| 0.0014*
Lower CL Dif 0.008415 Prob > t 0.0007*
Confidence 0.95 Prob < t 0.9993

These findings suggest that the upper to lower ratio of the 15th percentile of lung intensity may also be an independent predictor of lung cancer.

Considering these two significant variables into a single Generalized Linear Model (GLM) as independent variables and cancer as the outcome variable, i In this model, only the 15th percentile was significant (p=0.0055).

Generalized Linear Model Fit
Response: Cancer
Modeling P(Cancer=NO)
Distribution: Binomial
Link: Logit
Estimation Method: Maximum Likelihood

Observations (or Sum Wgts) = 210

Whole Model Test

Model	-LogLikelihood	L-R ChiSquare	DF	Prob>ChiSq
Difference	6.26033276	12.5207	2	0.0019*
Full	98.8241762			
Reduced	105.084509			

Effect Tests

Source	DF	L-R ChiSquare	Prob>ChiSq
R+L U/L mean	1	0.6205194	0.4309
R+L U/L 15%	1	7.7075925	0.0055*

Parameter Estimates

Term	Estimate	Std Error	L-R ChiSquare	Prob>ChiSq
Intercept	18.243874	6.0817368	10.530641	0.0012*
R+L U/L mean	3.6291227	4.568479	0.6205194	0.4309
R+L U/L 15%	-23.20951	8.6952667	7.7075925	0.0055*

This analysis suggests that only the upper to lower ratio of the 15th percentile of lung intensity may be a predictor of lung cancer.

KEY RESEARCH ACCOMPLISHMENTS:

- **Completion of sputum and plasma analysis from 210 subjects in a case control study of 150 with early stage lung cancer and 60 non cancer controls.**
- **Demonstration of specific and sensitive detection of cancer specific DNA methylation as an early detection biomarker.**
- **Transition of Studies to the University of Pittsburgh.**
- **Studies of emphysema and variability scores completed 127 subjects with a diagnosis of lung cancer and 180 subjects without a diagnosis of lung cancer.**

4. CONCLUSION:

In summary, based on our previous development of an improved panel of genes hypermethylated in lung cancer, with extraordinarily high specificity and sensitivity, we combined the improved methods of MOB with highly sensitive methylation specific PCR assays suitable for biologic fluid testing (sputum and serum) and completed the study of a cohort of cancer positive and negative samples. In combination with these molecular detection approaches, we have examined the alterations in air space for improving detection of lung cancer and find that variability of air spaces is associated with the presence of lung cancer. We have during the period of this grant developed a highly sensitive and specific method for early detection of lung cancer.

6. PUBLICATIONS, ABSTRACTS, AND PRESENTATIONS:

2015 Meeting of the American Association for Cancer Research (AACR) in Philadelphia, Pennsylvania
2015 Meeting of the International Association of Lung Cancer (IASLC) in Denver, Colorado

7. **INVENTIONS, PATENTS AND LICENSES:** Early Detection of Lung Cancer Using DNA Methylation in Plasma and Sputum, JHU C13599, licensing being pursued with Cepheid.
8. **REPORTABLE OUTCOMES:** Nothing to report
9. **OTHER ACHIEVEMENTS:** Nothing to report

10. REFERENCES: List all references pertinent to the report using a standard journal format (i.e. format used in Science, Military Medicine, etc.)

1. Wrangle J, Machida EO, Danilova L, Hulbert A, Franco N, Zhang W, Glockner SC, Tessema M, Van Neste L, Easwaran H, Schuebel KE, Licchesi J, Hooker CM, Ahuja N, Amano J, Belinsky SA, Baylin SB, Herman JG, Brock MV. Functional identification of cancer-specific methylation of CDO1, HOXA9, and TAC1 for the diagnosis of lung cancer. Clin Cancer Res. 2014;20(7):1856-64. doi: 10.1158/1078-0432.CCR-13-2109. PubMed PMID: 24486589; PubMed Central PMCID: PMC4019442.
2. Keeley B, Stark A, Pisanic TR, 2nd, Kwak R, Zhang Y, Wrangle J, Baylin S, Herman J, Ahuja N, Brock MV, Wang TH. Extraction and processing of circulating DNA from large sample volumes using methylation on beads for the detection of rare epigenetic events. Clinica chimica acta; international journal of clinical chemistry. 2013;425:169-75. doi: 10.1016/j.cca.2013.07.023. PubMed PMID: 23911908; PubMed Central PMCID: PMC3963364.
3. Howlader N, Noone A, Krapcho M, Garshell J, Miller D, Altekruse S, Kosary C, Yu M, Ruhl J, Tatalovich Z, Mariotto A, Lewis D, Chen H, Feuer E, Cronin K. SEER Cancer Statistics Review, 1975-2011 National Cancer Institute. Bethesda, MD, USA2014. Available from: http://seer.cancer.gov/csr/1975_2011/.
4. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2016. CA: a cancer journal for clinicians. 2016;66(1):7-30. doi: 10.3322/caac.21332. PubMed PMID: 26742998.
5. Jett JR. Current treatment of unresectable lung cancer. Mayo Clinic proceedings. 1993;68(6):603-11. Epub 1993/06/01. PubMed PMID: 8388526.
6. National Lung Screening Trial Research Team, Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, Gareen IF, Gatsonis C, Marcus PM, Sicks JD. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med. 2011;365(5):395-409. doi: 10.1056/NEJMoa1102873. PubMed PMID: 21714641; PubMed Central PMCID: PMC4356534.
7. Tammemagi MC, Katki HA, Hocking WG, Church TR, Caporaso N, Kvale PA, Chaturvedi AK, Silvestri GA, Riley TL, Commins J, Berg CD. Selection criteria for lung-cancer screening. N Engl J Med. 2013;368(8):728-36. doi: 10.1056/NEJMoa1211776. PubMed PMID: 23425165; PubMed Central PMCID: PMC3929969.
8. Bach PB, Mirkin JN, Oliver TK, Azzoli CG, Berry DA, Brawley OW, Byers T, Colditz GA, Gould MK, Jett JR, Sabichi AL, Smith-Bindman R, Wood DE, Qaseem A, Detterbeck FC. Benefits and harms of CT screening for lung cancer: a systematic review. JAMA. 2012;307(22):2418-29. doi: 10.1001/jama.2012.5521. PubMed PMID: 22610500; PubMed Central PMCID: PMC3709596.
9. Herman JG, Graff JR, Myohanen S, Nelkin BD, Baylin SB. Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. Proc Natl Acad Sci U S A. 1996;93(18):9821-6. doi: papers2://publication/uuid/8ED4493F-DFA1-485A-9048-572EC626410A. PubMed PMID: 8790415; PubMed Central PMCID: PMC38513.
10. Belinsky SA, Nikula KJ, Palmisano WA, Michels R, Saccomanno G, Gabrielson E, Baylin SB, Herman JG. Aberrant methylation of p16(INK4a) is an early event in lung cancer and a potential biomarker for early diagnosis. Proc Natl Acad Sci USA. 1998;95(20):11891-6. doi: Doi 10.1073/Pnas.95.20.11891. PubMed PMID: WOS:000076222200070.
11. Herman JG, Baylin SB. Gene silencing in cancer in association with promoter hypermethylation. N Engl J Med. 2003;349(21):2042-54. PubMed PMID: 14627790.
12. Belinsky SA. Gene-promoter hypermethylation as a biomarker in lung cancer. Nat Rev Cancer. 2004;4(9):707-17. PubMed PMID: 15343277.
13. Belinsky SA. Silencing of genes by promoter hypermethylation: key event in rodent and human lung cancer. Carcinogenesis. 2005;26(9):1481-7. doi: 10.1093/carcin/bgi020. PubMed PMID: 15661809.
14. Baylin SB, Ohm JE. Epigenetic gene silencing in cancer - a mechanism for early oncogenic pathway addiction? Nat Rev Cancer. 2006;6(2):107-16. PubMed PMID: 16491070.
15. Licchesi JD, Westra WH, Hooker CM, Herman JG. Promoter hypermethylation of hallmark cancer genes in atypical adenomatous hyperplasia of the lung. Clin Cancer Res. 2008;14(9):2570-8. PubMed PMID: 18451218.
16. Palmisano WA, Divine KK, Saccomanno G, Gilliland FD, Baylin SB, Herman JG, Belinsky SA. Predicting lung cancer by detecting aberrant promoter methylation in sputum. Cancer Res. 2000;60(21):5954-8. doi: papers2://publication/uuid/F0B0D370-349B-4CE3-9620-E76AC8B907CF. PubMed PMID: 11085511.
17. Belinsky SA, Liechty KC, Gentry FD, Wolf HJ, Rogers J, Vu K, Haney J, Kennedy TC, Hirsch FR, Miller Y, Franklin WA, Herman JG, Baylin SB, Bunn PA, Byers T. Promoter hypermethylation of multiple genes in sputum precedes lung cancer incidence in a high-risk cohort. Cancer Res. 2006;66(6):3338-44. PubMed PMID: 16540689.
18. Brock MV, Hooker CM, Ota-Machida E, Han Y, Guo M, Ames S, Glöckner S, Piantadosi S, Gabrielson E, Pridham G, Pelosky K, Belinsky SA, Yang SC, Baylin SB, Herman JG. DNA methylation markers and early recurrence in

stage I lung cancer. New England Journal of Medicine. 2008;358(11):1118-28. doi: papers2://publication/doi/10.1056/NEJMoa0706550.
19. Ostrow KL, Hoque MO, Loyo M, Brait M, Greenberg A, Siegfried JM, Grandis JR, Gaither Davis A, Bigbee WL, Rom W, Sidransky D. Molecular analysis of plasma DNA for the early detection of lung cancer by quantitative methylation-specific PCR. Clin Cancer Res. 2010;16(13):3463-72. doi: 10.1158/1078-0432.CCR-09-3304. PubMed PMID: 20592015; PubMed Central PMCID: PMC2899894.
20. Leng S, Do K, Yingling CM, Picchi MA, Wolf HJ, Kennedy TC, Feser WJ, Baron AE, Franklin WA, Brock MV, Herman JG, Baylin SB, Byers T, Stidley CA, Belinsky SA. Defining a gene promoter methylation signature in sputum for lung cancer risk assessment. Clin Cancer Res. 2012;18(12):3387-95. doi: 10.1158/1078-0432.CCR-11-3049. PubMed PMID: 22510351; PubMed Central PMCID: PMC3483793.
21. Li L, Shen Y, Wang M, Tang D, Luo Y, Jiao W, Wang Z, Yang R, Tian K. Identification of the methylation of p14ARF promoter as a novel non-invasive biomarker for early detection of lung cancer. Clin Transl Oncol. 2013;16(6):581-9. doi: papers2://publication/doi/10.1007/s12094-013-1122-1.
22. Sandoval J, Mendez-Gonzalez J, Nadal E, Chen G, Carmona FJ, Sayols S, Moran S, Heyn H, Vizoso M, Gomez A, Sanchez-Cespedes M, Assenov Y, Muller F, Bock C, Taron M, Mora J, Muscarella LA, Liloglou T, Davies M, Pollan M, Pajares MJ, Torre W, Montuenga LM, Brambilla E, Field JK, Roz L, Lo Iacono M, Scagliotti GV, Rosell R, Beer DG, Esteller M. A prognostic DNA methylation signature for stage I non-small-cell lung cancer. J Clin Oncol. 2013;31(32):4140-7. doi: 10.1200/JCO.2012.48.5516. PubMed PMID: 24081945.
23. Kim Y, Kim D-H. CpG Island Hypermethylation as a Biomarker for the Early Detection of Lung Cancer. (null). New York, NY: Springer New York; 2014. p. 141-71.
24. Nawaz I, Qiu X, Wu H, Li Y, Fan Y, Hu L-F, Zhou Q, Ernberg I. Development of a multiplex methylation specific PCR suitable for (early) detection of non-small cell lung cancer. Epigenetics. 2014;9(8):1138-48. doi: papers2://publication/doi/10.4161/epi.29499.
25. Yang X, Dai W, Kwong DL-w, Szeto CYY, Wong EH-w, Ng WT, Lee AWM, Ngan RKC, Yau CC, Tung SY, Lung ML. Epigenetic markers for noninvasive early detection of nasopharyngeal carcinoma by methylation-sensitive high resolution melting. Int J Cancer. 2014;136(4):E127-E35. doi: papers2://publication/doi/10.1002/ijc.29192.
26. Eads CA, Danenberg KD, Kawakami K, Saltz LB, Blake C, Shibata D, Danenberg PV, Laird PW. MethyLight: a high-throughput assay to measure DNA methylation. Nucleic Acids Res. 2000;28(8):E32. doi: papers2://publication/uuid/D055530D-DCBB-48AA-97F2-800F8D228527. PubMed PMID: 10734209; PubMed Central PMCID: PMC102836.
27. Bailey VJ, Zhang Y, Keeley BP, Yin C, Pelosky KL, Brock M, Baylin SB, Herman JG, Wang TH. Single-tube analysis of DNA methylation with silica superparamagnetic beads. Clin Chem. 2010;56(6):1022-5. PubMed PMID: 20360128.
28. Bailey VJ, Keeley BP, Razavi CR, Griffiths E, Carraway HE, Wang TH. DNA methylation detection using MS-qFRET, a quantum dot-based nanoassay. Methods. 2010;52(3):237-41. doi: 10.1016/j.ymeth.2010.03.007. PubMed PMID: 20362674.
29. Keeley B, Stark A, Pisanic TR, 2nd, Kwak R, Zhang Y, Wrangle J, Baylin S, Herman J, Ahuja N, Brock MV, Wang TH. Extraction and processing of circulating DNA from large sample volumes using methylation on beads for the detection of rare epigenetic events. Clin Chim Acta. 2013;425(C):169-75. doi: 10.1016/j.cca.2013.07.023. PubMed PMID: 23911908; PubMed Central PMCID: PMC3963364.
30. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012;489(7417):519-25. Epub 2012/09/11. doi: 10.1038/nature11404. PubMed PMID: 22960745; PubMed Central PMCID: PMC3466113.
31. Ettinger DS, Wood DE, Akerley W, Bazhenova LA, Borghaei H, Camidge DR, Cheney RT, Chirieac LR, D'Amico TA, Demmy TL, Dilling TJ, Govindan R, Grannis FW, Jr., Horn L, Jahan TM, Komaki R, Kris MG, Krug LM, Lackner RP, Lanuti M, Lilenbaum R, Lin J, Loo BW, Jr., Martins R, Otterson GA, Patel JD, Pisters KM, Reckamp K, Riely GJ, Rohren E, Schild S, Shapiro TA, Swanson SJ, Tauer K, Yang SC, Gregory K, Hughes M. NCCN Clinical Practice Guidelines in Oncology (NCCN Guidelines®) Non-small cell lung cancer, version 1.2015. J Natl Compr Canc Netw. 2014;12(12):1738-61. doi: papers2://publication/uuid/E03A4E8A-46FD-4107-8749-6F7122D9B452. PubMed PMID: 25505215.
32. Belinsky SA, Klinge DM, Dekker JD, Smith MW, Bocklage TJ, Gilliland FD, Crowell RE, Karp DD, Stidley CA, Picchi MA. Gene promoter methylation in plasma and sputum increases with lung cancer risk. Clin Cancer Res. 2005;11(18):6505-11. PubMed PMID: 16166426.
33. Prindiville SA, Byers T, Hirsch FR, Franklin WA, Miller YE, Vu KO, Wolf HJ, Baron AE, Shroyer KR, Zeng C, Kennedy TC, Bunn PA. Sputum cytological atypia as a predictor of incident lung cancer in a cohort of heavy smokers

with airflow obstruction. Cancer Epidemiol Biomarkers Prev. 2003;12(10):987-93. doi: papers2://publication/uuid/1DA01259-AC9B-4F24-BF1E-C75E39766623.
34. Genome Bioinformatics Group of UC Santa Cruz. UCSC Genome Bioinformatics 2015. Available from: http://genome.uscs.edu.
35. Brandes JC, Carraway H, Herman JG. Optimal primer design using the novel primer design program: MSPprimer provides accurate methylation analysis of the ATM promoter. Oncogene. 2007;26(42):6229-37. PubMed PMID: 17384671.
36. Untergrasser A CI, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG Primer3web 2012. Available from: http://primer3.ut.ee/.
37. Team RC. R: A language and environment for statistical computing. R-Project Org, Version 3.0.2 ed. Vienna, Austria: R Foundation for Statistical Computing; 2013.
38. Etzioni R, Urban N, Ramsey S, McIntosh M, Schwartz S, Reid B, Radich J, Anderson G, Hartwell L. The case for early detection. Nat Rev Cancer. 2003;3(4):243-52. Epub 2003/04/03. doi: 10.1038/nrc1041. PubMed PMID: 12671663.
39. Kennedy TC, Proudfoot SP, Piantadosi S, Wu L, Saccomanno G, Petty TL, Tockman MS. Efficacy of two sputum collection techniques in patients with air flow obstruction. Acta Cytol. 1999;43(4):630-6. Epub 1999/08/05. PubMed PMID: 10432886.
40. Kennedy TC, Proudfoot SP, Franklin WA, Merrick TA, Saccomanno G, Corkill ME, Mumma DL, Sirgi KE, Miller YE, Archer PG, Prochazka A. Cytopathological analysis of sputum in patients with airflow obstruction and significant smoking histories. Cancer Res. 1996;56(20):4673-8. doi: papers2://publication/uuid/270AD142-0700-4EB0-A6E1-662170FFC855. PubMed PMID: 8840983.
41. Silvestri GA, Vachani A, Whitney D, Elashoff M, Porta Smith K, Ferguson JS, Parsons E, Mitra N, Brody J, Lenburg ME, Spira A, Team AS. A Bronchial Genomic Classifier for the Diagnostic Evaluation of Lung Cancer. N Engl J Med. 2015;373(3):243-51. Epub 2015/05/20. doi: 10.1056/NEJMoa1504601. PubMed PMID: 25981554.

11. APPENDICES: Manuscript in pages to follow

EARLY DETECTION OF LUNG CANCER USING DNA PROMOTER HYPERMETHYLATION IN PLASMA AND SPUTUM

Alicia Hulbert,[1,2*] Ignacio Jusue-Torres,[3*] Alejandro Stark,[4*] Chen Chen,[1,5*] Kristen Rodgers,[2] Beverly Lee,[2] Candace Griffin,[2] Andrew Yang,[2] Peng Huang,[1, 6] John Wrangle,[7] Steven A Belinsky,[8] Tza-Huei Wang,[1,4,9] Stephen C Yang,[2] Stephen B Baylin,[1] Malcolm V Brock,[1,2] James G Herman.[1,10]

[1] Sidney Kimmel Cancer Center. Department of Oncology, The Johns Hopkins University. School of Medicine, Baltimore, MD.
[2] Department of Surgery. The Johns Hopkins University School of Medicine, Baltimore, MD
[3] Department of Neurosurgery. The Johns Hopkins University School of Medicine, Baltimore, MD
[4] Department of Mechanical Engineering. The Johns Hopkins University, Baltimore, MD
[5] Department of Thoracic Surgery, Second Xiangya Hospital of Central South University, Changsha, Hunan, P.R China
[6] Department of Biostatistics, The Johns Hopkins Bloomberg School of Public Health, Baltimore, MD
[7] Department of Medicine, Medical University of South Caroline, Charleston, SC.
[8] Lung Cancer Program, Lovelace Respiratory Research Institute, Albuquerque, New Mexico
[9] Department of Biomedical Engineering and Institute for NanoBioTechnology. The Johns Hopkins University. School of Medicine, Baltimore, MD.
[10] Department of Medicine, University of Pittsburgh School of Medicine, Pittsburg, PA.

Acknowledgments:
*****contributed equally.
Funding: DOD W81XWH-12-1-0323 and SPORE P50CA058184.

Corresponding author:
James G. Herman
Professor of Medicine
Director of the Lung Cancer Program
University of Pittsburgh, Division of Hematology/ Oncology
2.18d Hillman Cancer Center
Pittsburgh, PA 15232
Email: hermanj3@upmc.edu
Tel: 412-623-7769
Fax: 412-623-7768

Running head:
Epigenetic Lung cancer screening

The abstract of this manuscript has been presented at:
2015 Meeting of the American Association for Cancer Research (AACR) in Philadelphia, Pennsylvania
2015 Meeting of the International Association of Lung Cancer (IASLC) in Denver, Colorado

Early Detection of Lung Cancer using DNA Promoter Hypermethylation in Plasma and Sputum

Abstract:

Purpose[AH1]**:**

To improve the diagnostic accuracy of lung cancer screening using ultrasensitive methods detecting gene promoter methylation in sputum and plasma using Methylation-On-Beads (MOB) with a lung cancer specific gene panel.

Patients and Methods:

This is a case-control study of subjects with nodules suspicious for lung cancer on CT imaging in which plasma and sputum were obtained pre-operatively. Cases (n=150) had pathological confirmation of node negative (stage IA, IB and IIA) non-small cell lung cancer while controls (n=60) had non-cancer diagnoses. We detected promoter methylation using quantitative methylation specific real-time PCR with MOB for cancer-specific genes (CDO1, TAC1, HOXA7, HOXA9, SOX17 and ZFP42) identified from The Cancer Genome Atlas (TCGA).

Results:

DNA methylation was detected in plasma and sputum more frequently in people with cancer compared to controls ($p<0.001$) for 5 of 6 genes examined. Individual gene detection The sensitivity and specificity for lung cancer diagnosis using individual genes from sputum ranged from 63-93% and 42-92% respectively and from plasma from 33-91% and 52-94%. A three-gene combination including the best individual genes has sensitivity and specificity of 93% and 79% using sputum and 91% and 64% using plasma. Area under the Receiver Operating Curve for this panel was 0.89 95% CI (0.80-0.98) in sputum and 0.77 95% CI (0.68-0.86) in plasma. Independent, blinded random forest prediction models combining gene methylation with age, pack-year, COPD status and FVC values correctly predicted lung cancer in 91% of subjects using sputum samples and 85% of subjects using blood samples.

Conclusions:

High diagnostic accuracy for early stage lung cancer can be obtained using methylated promoter detection in sputum or plasma.

Keywords:

Lung cancer screening; epigenetics; gene promoter hypermethylation; early detection; molecular biomarkers

Background

Lung cancer is the third most prevalent cancer with over 224,000 [HJG2]cases annually in the U.S.(3, 4) It is the most deadly cancer worldwide accounting for almost 27% of all cancer-related deaths in part because of advanced stage at diagnosis in 67% of cases.(3, 5) The National Lung Screening Trial (NLST) demonstrated a 20% reduction in lung cancer mortality using low-dose computed tomography (CT) screening.[(6)] This survival benefit comes at the price of detecting many indeterminate, small pulmonary nodules with a false positive rate of 96.4%.(6, 7) This has led to cautious adoption of CT screening, because complications, and even deaths, result from further diagnostic procedures.(8)

One approach to improving the specificity of CT screening involves the use of cancer specific biomarkers from sputum and plasma. The epigenetic alteration of promoter DNA methylation is associated with the initiation and progression of cancer,(9-15) and may be used as a biomarker for cancer risk, prevention, treatment, and prognosis(1, 16-25) However previous approaches had limited sensitivity and specificity and were not adequate for lung cancer screening.(16-25)

Reduced sensitivity of methylation detection may occur from technical limitations. Traditional extraction methods for DNA, such as phenol-chloroform, are inefficient for extracting small amounts of DNA due to repeated sample transfers with sample loss and degradation of DNA during bisulfite conversion.(9, 26) We have developed Methylation-on-Beads (MOB) which successfully combines these processes into a single process, reducing sample loss with potentially increased sensitivity.(27-29)

In addition, previous studies have selected genes for DNA methylation detection primarily chosen from a candidate approach, which are methylated in only a fraction of tumors. The Cancer Genome Atlas (TCGA)(30) provides the opportunity to discover cancer specific methylation changes optimal for detection. We had . reported the identification of six genes (CDO1, HOXA7, HOXA9, TAC1, SOX17, and ZFP42) with a high prevalence of methylation changes present in lung squamous and adenocarcinoma, but not normal lung tissue.(1) These were developed into sensitive assays using MOB and real-time Methylation-Specific PCR (qMSP) to determine the diagnostic accuracy in sputum and plasma for lung cancer detection in a case-control study.

Patients and Methods

Study Population

The study population consists of a prospective, observational cohort of 651 participants, initiated in 2007 within the Johns Hopkins Lung Cancer Specialized Program of Research Excellence (SPORE), to monitor cancer recurrence after surgery. From this cohort, 210 study patients had node negative early stage tumors (T1-T2N0) and samples adequate for analysis. Institutional review board approval was obtained prior to the start of this study (NA_00005998), and all patients signed informed consent. Surgical resection with curative intent and pathological analyses of suspected lung cancer lesions were completed in all patients. Patients were staged according to the new revised TNM guidelines classification criteria.(31) Cases were defined as patients with confirmed lung cancer by pathology. Controls were defined as patients histologically confirmed not to have cancer. Plasma and sputum samples were obtained prior to surgical resection. Pack-years of cigarette smoking was defined as the average number of packs smoked per day multiplied by the number of years of smoking. Nodule size was obtained from the pathological report. Nodule volume, obtained from surgical pathological reports, was calculated with the ellipsoid volume formula (Volume = $4/3 \times \pi \times$ radius A $\times$ radius B $\times$ radius C).

Plasma and Sputum Collection

Prior to surgery, 20 ml of plasma was collected in tubes containing sodium heparin (Bectin Dickinson, Franklin Lakes) and then stored at -80°C. For sputum collection, two cups containing Saccomanno's fixative solution were used for each patient as previously described.(17, 20, 32) Subjects were asked to provide an early morning spontaneous sputum at home in two cups for 3 consecutive days within 1 week prior to pulmonary resection.(20, 33) Five milliliters of sputum was collected, washed with Saccomanos' solution, vortexed, centrifuged and then stored at -80ºC.(17)

DNA Isolation and Bisulfite Conversion

DNA extraction from tumor, plasma and sputum was performed using MOB, a process that allows DNA extraction and bisulfite conversion in a single tube via the use of silica super magnetic beads.(27) This approach yields a 1.5 to 5-fold improvement in extraction efficiency with a small

amount of DNA in comparison to traditional conventional techniques.(29) We have optimized the protocol previously described for plasma(29), using 1.5 ml of plasma and 375 ul (800units/ml, NEBL p8107s) of proteinase K. For DNA extraction from sputum using the MOB method, we modified the protocol used for plasma by adding 200 ul of sample to 300 ul of Buffer AL and 40 ul of Proteinase K and by incubating them together at the same temperature (50 °C for 2 hours). After digestion, 300 ul of IPA and 150 ul of beads were added. The lysate was also incubated and rotated for 10 minutes before adding 5 ul of carrier RNA, and incubating for an additional 5 minutes.(29)

DNA Methylation Analysis

The genomic sequence for the genes and 1000 bases upstream was obtained from the UCSC genomic browser website.(34) The primers and hybridization probes for methylation analysis were designed based on this sequence by using Primer3 (v.0.4.0).(35, 36) All primer and probe sequences are listed in supplementary **Table S1**. The analysis was performed using quantitative real-time Methylation Specific PCR, and normalized to a control β-Actin assay.(26) Each reaction was performed in a 25 µl PCR mixture consisting of 2 µl of bisulfite converted DNA, 300 nM R-sense primer, 300 nM F-anti-sense primer, 100nM probe, 100 nM of fluorescein reference dye (Life Technologies), 1.67mM dNTPs (VWRQuotation), and 1 ul of Platinum Taq® DNA Polymerase (invitrogen). Master mix contained 16.6mM $(NH_4)_2SO_4$, 67mM Tris pH 8.8, 6.7 mM $MgCl_2$ and 10mM β-mercaptoethanol in a nuclease-free DI water solution. Amplification reactions were performed using 96 well-plates (MicroAmp®) with all samples being analyzed in triplicate. Thermo cycling conditions were as follows: 95°C for 5 min, 50 cycles at 95°C for 15 seconds, and 65°C for 1min and 72°C for 1 min. An ABI StepOnePlus Real-Time PCR system was used (Applied Bio Systems, examples shown in **Supplemental Figure 1**).

With the extremely low levels of DNA methylation in plasma and sputum, replicates for some samples produced no detectable methylation as expected. To incorporate this information into the final quantification of methylation, we calculated the $2^{-\Delta CT}$ for each methylation detection replicate comparing it to the mean Ct for β-Actin (ACTB). For replicates which were not detected (ND), a CT of 100 was used, creating a near zero value for $2^{-\Delta CT}$. The mean $2^{-\Delta CT}$ value was calculated with the formula:

$$\mu\, 2^{-\Delta CT} = \frac{(2^{-\Delta CT replicate\ 1} + 2^{-\Delta CT replicate\ 2} + 2^{-\Delta CT replicate\ 3})}{3}$$

Statistical Analysis

Quantitative data are expressed as median (interquartile range) for continuous, non-parametric variables and frequency (percentage) for categorical variables. For inter-group comparison, the Wilcoxon rank sum test was used for continuous data and the Fisher´s exact test for categorical data[HJG3].

Data was analyzed using two approaches. The first approach is the ROC analysis using the $2^{-\Delta CT}$ values for individual genes to determine the performance of each individual marker. The three best performing genes were selected for diagnostic accuracy for lung cancer detection, based on receiver operator classification (ROC) curves and were used for combined detection. Sensitivity and specificity values were obtained from the optimum cutoff thresholds from ROC curves (R statistic software, version 3.0.2, Vienna, Austria).[37] The area under the curve was reported with 95% CIs.

The second approach utilized independent blinded random forest prediction models, a non-parametric machine learning method, to evaluate the utility of the six-gene panel and clinical data in early lung cancer detection. The analysis combined gene methylation with clinical risk factors: nodule size, age, pack-year, COPD status and FVC values (**Figure 4**). Two-thirds of subjects were randomly selected as a training set and the remainder formed the test set. A statistician (PH), blinded to the true diagnosis codes of the test set patients, used the training set to build three random forest prediction models: the first one used all six-gene sputum biomarkers plus clinical and demographic risk factors, the second used the six-gene plasma biomarkers and clinical, and a third used only clinical and demographic risk factors without any methylation biomarkers. These three models were ysed to predict cancer status from the independent test set. Prediction accuracy was reported as the proportion of test set subjects correctly predicted by the random forest classification models, allowing calculation of sensitivity, specificity, and ROC analysis.

Results

Characteristics of the Patients

Two hundred and ten patients fulfilled inclusion criteria, , with 150 node negative early stage

lung cancer subjects and 60 controls with non-cancerous lung lesions (**Table 1**). Clinical and demographic variables were similar in cases and controls with the exception of age, number of pack-year and nodule size (cm) as well as volume (cm^3). Subjects with lung cancer were significantly older than controls (67 vs. 73 years, p=0.007), smoked significantly more (30 vs. 19.5 pack years, p=0.01), and had significantly larger nodules (2.0 vs. 1.5 cm, p=0.01). The proportion of smokers, former smokers and never smokers was not different between cases and controls.

Detection of DNA Methylation

We measured DNA methylation for these genes in tumor tissue, confirming our previous study suggesting these genes were methylated in the majority of lung tumors (Figure 1). Methylation in sputum was detected more frequently in all 6 genes in cancer patients compared to controls (**Figure 1**[J4]), which for some patients was quantitatively similar to lung tumor tissues, but in some cases was at levels previously below conventional methods of detection. For 5 of the 6 genes, (CDO1, TAC1, HOXA7, SOX17 and ZFP42) this was statistically significant ($p < 0.001$). Methylation of all 6 genes was detected more frequently in plasma in cases compared to controls ($p < 0.001$). The worst performing gene was HOXA9 in plasma, which showed a lack of specificity as was also seen in the sputum**.** We determined the sensitivity and specificity in this cohort using the presence or absence of detectable methylation without considering the quantitation of methylation. This resulted in good sensitivity and specificities (Table 2a).

Gene Methylation and Lung Cancer Diagnostic Accuracy[J5]

ROC curves for lung cancer detection were [AH6]obtained for each single gene; using the normalized methylation ΔC_t values calculated as described in methods (**Table 2**, ROC curves in **Supplemental figure 2 & 3**). The sensitivity and specificity for lung cancer diagnosis from single methylated genes in sputum ranged 63-93% and 42-92% respectively and in plasma from 33-91% and 52-94%. The AUC[HJ7] values were 0.56-0.89 in sputum samples and 0.60-0.78 in plasma samples.

The genes with the largest AUC in sputum were: **TAC1** AUC: 0.84 95% CI (0.74-0.94), **SOX17** AUC: 0.84 95% CI (0.75-0.94) and **HOXA7** AUC: 0.77 95% CI (0.67-0.86) (**Figure 2A)**, with sensitivities and specificities for TAC1 84% and 79%; SOX17 84% and 88%; HOXA7 63% and 92% respectively. The positive and negative predictive values for these three genes were: TAC1 94% and

57%; SOX17 96% and 59%; HOXA7 97% and 40% respectively.

In plasma, the genes with the largest areas under the curve (AUC) were: **CDO1** AUC: 0.68 95% CI (0.58-0.77), **TAC1** AUC: 0.78 (0.70-0.86) and **SOX17** AUC: 0.78 95% CI (0.70-0.86) (**Figure 2B**), with corresponding sensitivities and specificities of: CDO1 65% and 74%; TAC1 76% and 78%; SOX17 71% and 86% respectively. The positive and negative predictive values for these genes were: CDO1 86% and 46%; TAC1 90% and 57%; SOX17 93% and 54% respectively.

The sensitivity and specificity derived from the optimum cutoff point obtained from the ROC curve in the combination of the three best performing markers (TAC1, SOX17 and HOXA7) in sputum was 93% and 79%, respectively with a corresponding ROC AUC of 0.89 95% CI (0.80-0.98) (**Figure 2C**). In plasma, the combination of CDO1, TAC1 and SOX17 showed a sensitivity, specificity and AUC of 91%, 64% and 0.77, 95% CI (0.68-0.86). respectively (**Figure 2D**).

Smokers subset analysis

Since CT screening for lung cancer is currently recommended for current and ex-smokers, we explored the diagnostic accuracy when only smokers were considered (n=155; 114 with cancer and 41 without cancer) (**Supplemental Tables for Only Smokers S2**). The results in only smokers were similar to the entire study population for the prevalence of methylated patients, sensitivity, specificity and AUC (**Supplemental Table S3**). AUC in smokers only was 0.89 95% CI (0.79-0.99) for the combination of the methylation status of the best three genes from sputum and AUC 0.85 95% CI (0.76-0.94) from the best three genes from plasma (**Supplemental Table S4 & S5**).

Independent Prediction Accuracy Performance

While the above analysis looked at individual gene methylation in cases and controls to detect cancer, independent blinded random forest prediction models analyzed all these biomarkers in combination with clinical risk factors. Risk factors included in the first two random forest prediction models were methylation *C*t values from all six genes, age, pack-year, COPD status and FVC values. The methylation *C*t values were not included in the last prediction model. The randomly selected **training** dataset has 140 subjects with 99 (70.7%) cancers and 41 (29.3%) controls. The independent **test** set has 70 subjects with 51 (72.9%) cancers and 19 (27.1%) controls. In the variable of importance output of the first two random forest prediction models, methylation *C*t values were ranked as more

important variables than demographic and clinical variables (**Figure 3 and 4**). **Table 3** summarizes the prediction accuracies of these three[HJ8] models when they were applied to the independent test set patients. With **sputum samples**, the random forest model correctly predicted lung cancer in 91% of subjects in the test subset. The corresponding AUC was 0.85 95% CI (0.59-1.0) **(Figure 3)**. The sensitivity and specificity of the prediction in the testing subset from the ROC curve were 0.93 and 0.86, respectively. Using **plasma samples**, the random forest model correctly predicted lung cancer in 85% of subjects in the testing subset. The corresponding AUC was 0.89 95% CI (0.79-0.99) **(Figure 3)**. The sensitivity and specificity of the prediction in the testing subset from the ROC curve were 0.93 and 0.67, respectively. Using clinical and demographic risk factors alone, the accuracies were lower than the first two models with a diagnostic accuracy of 68%, AUC of 0.64, PPV of 75% and a NPV of 38% (**Figure 3 and Table 3)**.

Discussion

High diagnostic accuracy for early stage lung cancer can be obtained using a panel of methylated promoter genes in sputum or plasma, and an ultrasensitive detection strategy based on MOB. This assay has several characteristics which make it clinically useful (i) it has a sensitivity and specificity in sputum and plasma which exceeds the diagnostic accuracy required by most clinical standards(12, 38) (ii) it can be performed with minute quantities of DNA from sputum or plasma (iii) it can be used to distinguish malignant versus benign CT detected nodules, addressing the current problem of high false positive CT findings in lung cancer screening. This discrimination is associated with risk of lung cancer independent of age, pack-year and nodule size, and is can detect early stage lung cancer in smokers. Finally, as a PCR-based assay, it is simple and relatively inexpensive.

Previous studies have sought to improve lung cancer risk assessment by the use of molecular biomarkers obtained from blood and sputum.(17, 19, 20, 32, 33, 39, 40) However none of these tests have been used clinically because their achieved sensitivities and specificities were usually not high enough for clinical decision-making.(17, 19, 20, 32, 33, 39-41) With improvements in DNA extraction methods and processing for methylation detection, along with the use of highly prevalent cancer specific methylation targets, we have overcome these obstacles.

In this study, detection of methylation in sputum samples was slightly better than the detection of these same genes in plasma. The access of early cancers to the airways may be one explanation for

this difference. Indeed, changes in the airways form the basis for the AEGIS Study, which reported an improved diagnostic yield of bronchoscopy using gene-expression classifiers from epithelial cells collected during bronchoscopy.[41] The AUC, sensitivities and specificities reported in the AEGIS Study were lower than the ones in the present study.

In our model where methylation markers from blood were considered simultaneously with age and number of pack-years, we observed a predictive accuracy close to that of sputum. This suggests that blood could substitute for sputum in lung cancer detection in those cases where sputum cannot be obtained.

According to the NLST, the chances of having lung cancer with a positive CT screening are less than 5%.(6, 7) This is because lung cancer with CT screening in the NLST study yielded a 71% sensitivity but a 63% specificity with a 96.4% false positive rate.(6, 7) Our current findings indicate that methylation detection using a few genes from blood and sputum could potentially reduce false positive screening. Although our study included patients who would not meet current lung cancer screening guidelines, we observed similar detection rates when only smokers were analyzed. Replication and external validation of our findings in a large, prospective, multicenter case control trial are essential before this approach can be adopted.

Conclusion

This study shows that it is possible to obtain high sensitivity and specificity detection of early stage NSCLC using a panel of methylated promoter genes in plasma and sputum, and that the methylation level of these genes is associated with a high lung cancer risk independent of age, pack-year and nodule size. These epigenetic biomarkers could be used as an adjunct to CT screening to identify patients at high risk for lung cancer, reducing false positive results, unnecessary tests, as well as improving the diagnosis of lung cancer at an earlier stage.

References

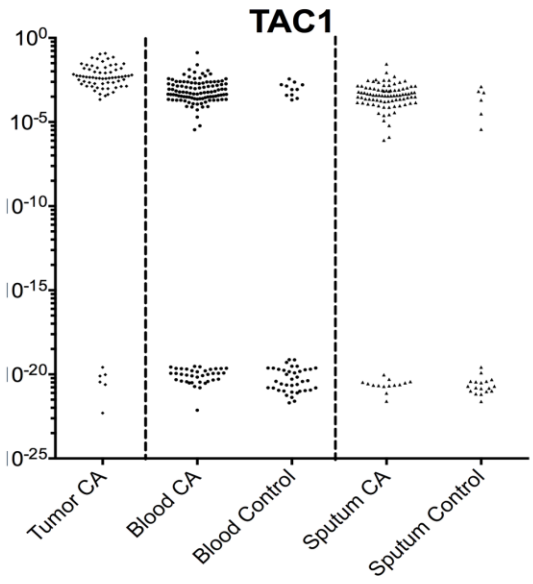
1. Howlader N, Noone A, Krapcho M, *et al. SEER Cancer Statistics Review, 1975-2011*. http://seer.cancer.gov/csr/1975_2011/.
2. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2016. CA Cancer J Clin 2016;66(1):7-30.
3. Jett JR. Current treatment of unresectable lung cancer. Mayo Clin Proc 1993;68(6):603-11.
4. National Lung Screening Trial Research Team, Aberle DR, Adams AM, *et al.* Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 2011;365(5):395-409.
5. Tammemagi MC, Katki HA, Hocking WG, *et al.* Selection criteria for lung-cancer screening. N Engl J Med 2013;368(8):728-36.
6. Bach PB, Mirkin JN, Oliver TK, *et al.* Benefits and harms of CT screening for lung cancer: a systematic review. JAMA 2012;307(22):2418-29.
7. Herman JG, Graff JR, Myohanen S, *et al.* Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. Proc Natl Acad Sci U S A 1996;93(18):9821-6.
8. Belinsky SA, Nikula KJ, Palmisano WA, *et al.* Aberrant methylation of p16(INK4a) is an early event in lung cancer and a potential biomarker for early diagnosis. Proceedings of the National Academy of Sciences of the United States of America 1998;95(20):11891-11896.
9. Herman JG, Baylin SB. Gene silencing in cancer in association with promoter hypermethylation. N Engl J Med 2003;349(21):2042-54.
10. Belinsky SA. Gene-promoter hypermethylation as a biomarker in lung cancer. Nat Rev Cancer 2004;4(9):707-17.
11. Belinsky SA. Silencing of genes by promoter hypermethylation: key event in rodent and human lung cancer. Carcinogenesis 2005;26(9):1481-7.
12. Baylin SB, Ohm JE. Epigenetic gene silencing in cancer - a mechanism for early oncogenic pathway addiction? Nat Rev Cancer 2006;6(2):107-16.
13. Licchesi JD, Westra WH, Hooker CM, *et al.* Promoter hypermethylation of hallmark cancer genes in atypical adenomatous hyperplasia of the lung. Clin Cancer Res 2008;14(9):2570-8.
14. Palmisano WA, Divine KK, Saccomanno G, *et al.* Predicting lung cancer by detecting aberrant promoter methylation in sputum. Cancer Res 2000;60(21):5954-8.
15. Belinsky SA, Liechty KC, Gentry FD, *et al.* Promoter hypermethylation of multiple genes in sputum precedes lung cancer incidence in a high-risk cohort. Cancer Res 2006;66(6):3338-44.
16. Brock MV, Hooker CM, Ota-Machida E, *et al.* DNA methylation markers and early recurrence in stage I lung cancer. The New England journal of medicine 2008;358(11):1118-1128.
17. Ostrow KL, Hoque MO, Loyo M, *et al.* Molecular analysis of plasma DNA for the early detection of lung cancer by quantitative methylation-specific PCR. Clin Cancer Res 2010;16(13):3463-72.
18. Leng S, Do K, Yingling CM, *et al.* Defining a gene promoter methylation signature in sputum for lung cancer risk assessment. Clin Cancer Res 2012;18(12):3387-95.
19. Li L, Shen Y, Wang M, *et al.* Identification of the methylation of p14ARF promoter as a novel non-invasive biomarker for early detection of lung cancer. Clinical & translational oncology : official publication of the Federation of Spanish Oncology Societies and of the National Cancer Institute of Mexico 2013;16(6):581-589.
20. Sandoval J, Mendez-Gonzalez J, Nadal E, *et al.* A prognostic DNA methylation signature for stage I non-small-cell lung cancer. J Clin Oncol 2013;31(32):4140-7.
21. Kim Y, Kim D-H. CpG Island Hypermethylation as a Biomarker for the Early Detection of Lung Cancer. In. *(null)*. New York, NY: Springer New York; 2014, 141-171.
22. Nawaz I, Qiu X, Wu H, *et al.* Development of a multiplex methylation specific PCR suitable for (early) detection of non-small cell lung cancer. Epigenetics 2014;9(8):1138-1148.
23. Wrangle J, Machida EO, Danilova L, *et al.* Functional identification of cancer-specific methylation of CDO1, HOXA9, and TAC1 for the diagnosis of lung cancer. Clin Cancer Res 2014;20(7):1856-64.

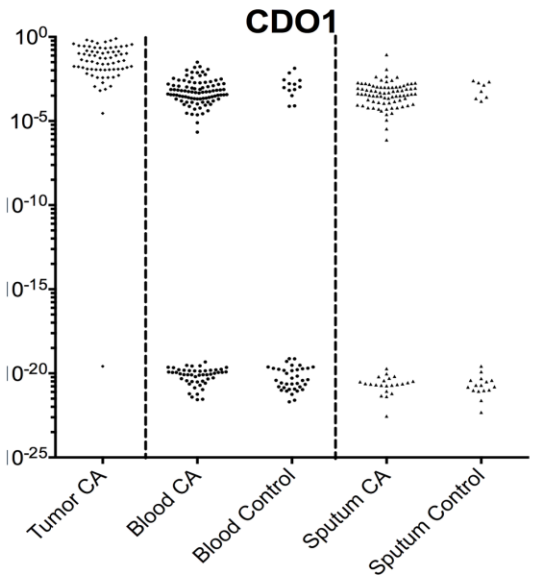
24. Yang X, Dai W, Kwong DL-w, *et al.* Epigenetic markers for noninvasive early detection of nasopharyngeal carcinoma by methylation-sensitive high resolution melting. International Journal of Cancer 2014;136(4):E127-E135.
25. Eads CA, Danenberg KD, Kawakami K, *et al.* MethyLight: a high-throughput assay to measure DNA methylation. Nucleic Acids Res 2000;28(8):E32.
26. Bailey VJ, Zhang Y, Keeley BP, *et al.* Single-tube analysis of DNA methylation with silica superparamagnetic beads. Clin Chem 2010;56(6):1022-5.
27. Bailey VJ, Keeley BP, Razavi CR, *et al.* DNA methylation detection using MS-qFRET, a quantum dot-based nanoassay. Methods 2010;52(3):237-41.
28. Keeley B, Stark A, Pisanic TR, 2nd, *et al.* Extraction and processing of circulating DNA from large sample volumes using methylation on beads for the detection of rare epigenetic events. Clin Chim Acta 2013;425(C):169-75.
29. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. Nature 2012;489(7417):519-25.
30. Ettinger DS, Wood DE, Akerley W, *et al.* NCCN Clinical Practice Guidelines in Oncology (NCCN Guidelines®) Non-small cell lung cancer, version 1.2015. J Natl Compr Canc Netw 2014;12(12):1738-61.
31. Belinsky SA, Klinge DM, Dekker JD, *et al.* Gene promoter methylation in plasma and sputum increases with lung cancer risk. Clin Cancer Res 2005;11(18):6505-11.
32. Prindiville SA, Byers T, Hirsch FR, *et al.* Sputum cytological atypia as a predictor of incident lung cancer in a cohort of heavy smokers with airflow obstruction. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology 2003;12(10):987-993.
33. Genome Bioinformatics Group of UC Santa Cruz. *UCSC Genome Bioinformatics*. http:// genome.uscs.edu.
34. Brandes JC, Carraway H, Herman JG. Optimal primer design using the novel primer design program: MSPprimer provides accurate methylation analysis of the ATM promoter. Oncogene 2007;26(42):6229-37.
35. Untergrasser A CI, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG *Primer3web*. http://primer3.ut.ee/.
36. Team RC. R: A language and environment for statistical computing. In. R-Project Org, Version 3.0.2 ed. Vienna, Austria: R Foundation for Statistical Computing; 2013.
37. Etzioni R, Urban N, Ramsey S, *et al.* The case for early detection. Nat Rev Cancer 2003;3(4):243-52.
38. Kennedy TC, Proudfoot SP, Piantadosi S, *et al.* Efficacy of two sputum collection techniques in patients with air flow obstruction. Acta Cytol 1999;43(4):630-6.
39. Kennedy TC, Proudfoot SP, Franklin WA, *et al.* Cytopathological analysis of sputum in patients with airflow obstruction and significant smoking histories. Cancer Res 1996;56(20):4673-8.
40. Silvestri GA, Vachani A, Whitney D, *et al.* A Bronchial Genomic Classifier for the Diagnostic Evaluation of Lung Cancer. N Engl J Med 2015;373(3):243-51.

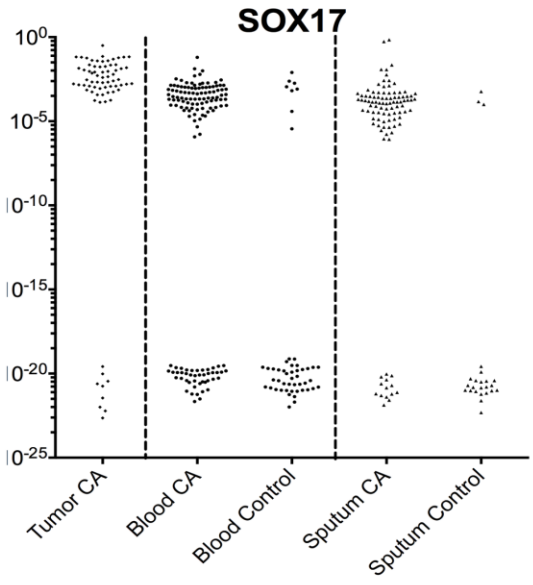
Figures

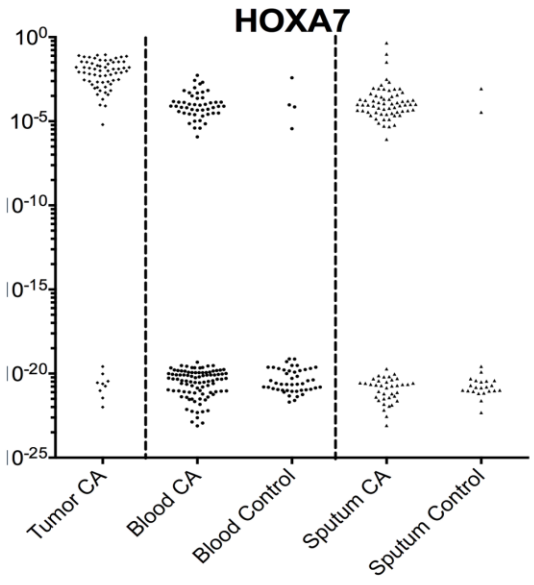
Figure 1. Methylation detection values of the studied genes.

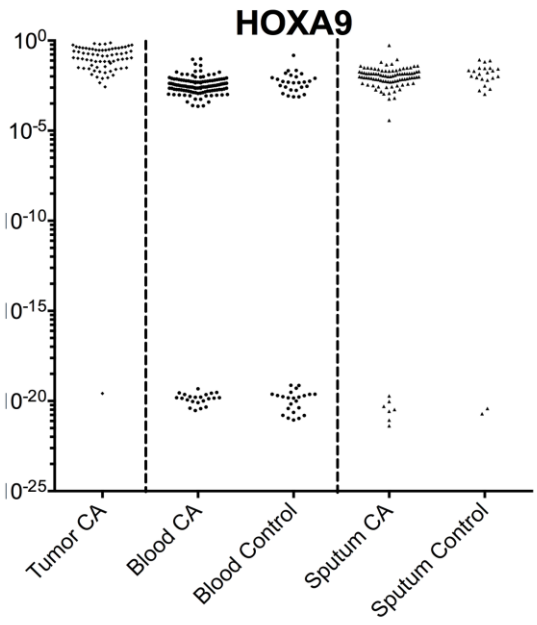
This scatter plot shows the converted ΔCT methylation values in a logarithmic scale. These values show a bimodal distribution with the lower group the values corresponding to those samples with no detectable amplification (ND). The majority of lung tumor samples have high levels of methylation, as expected from the previous study. Blood and sputum samples from cancer patients have detectable methylation which varies from levels nearing that of tumor samples to those at the limits of detection (10^{-5}-10^{-6}), while some patients are undetectable. The majority of controls have undetectable methylation at these loci, although some patients do have detectable methylation that is quantitatively similar to cancer patients. HOXA9 methylation is detectable in most control patients, especially in the sputum, suggesting this change is present in the lung epithelium and not as specific for the detection of cancer[HJG9].


TAC1
10^{0}
10^{-5}
10^{-10}
10^{-15}
10^{-20}
10^{-25}
Tumor CA
Blood CA
Blood Control
Sputum CA
Sputum Control


CDO1
10^{0}
10^{-5}
10^{-10}
10^{-15}
10^{-20}
10^{-25}
Tumor CA
Blood CA
Blood Control
Sputum CA
Sputum Control


SOX17
10^{0}
10^{-5}
10^{-10}
10^{-15}
10^{-20}
10^{-25}
Tumor CA
Blood CA
Blood Control
Sputum CA
Sputum Control


HOXA7
10^{0}
10^{-5}
10^{-10}
10^{-15}
10^{-20}
10^{-25}
Tumor CA
Blood CA
Blood Control
Sputum CA
Sputum Control


HOXA9
10^{0}
10^{-5}
10^{-10}
10^{-15}
10^{-20}
10^{-25}
Tumor CA
Blood CA
Blood Control
Sputum CA
Sputum Control

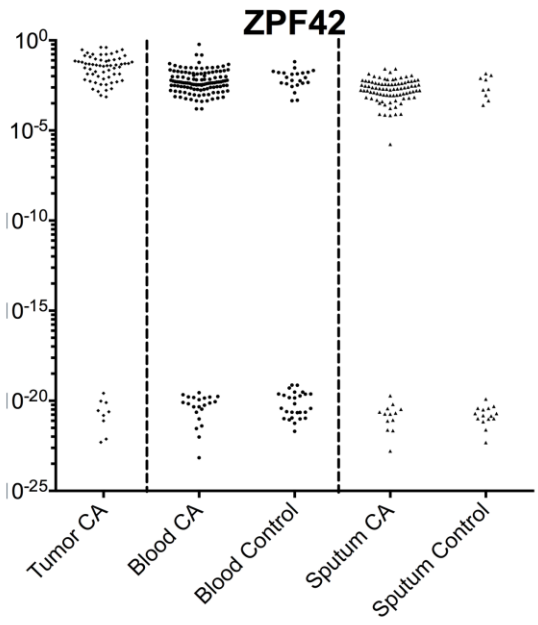
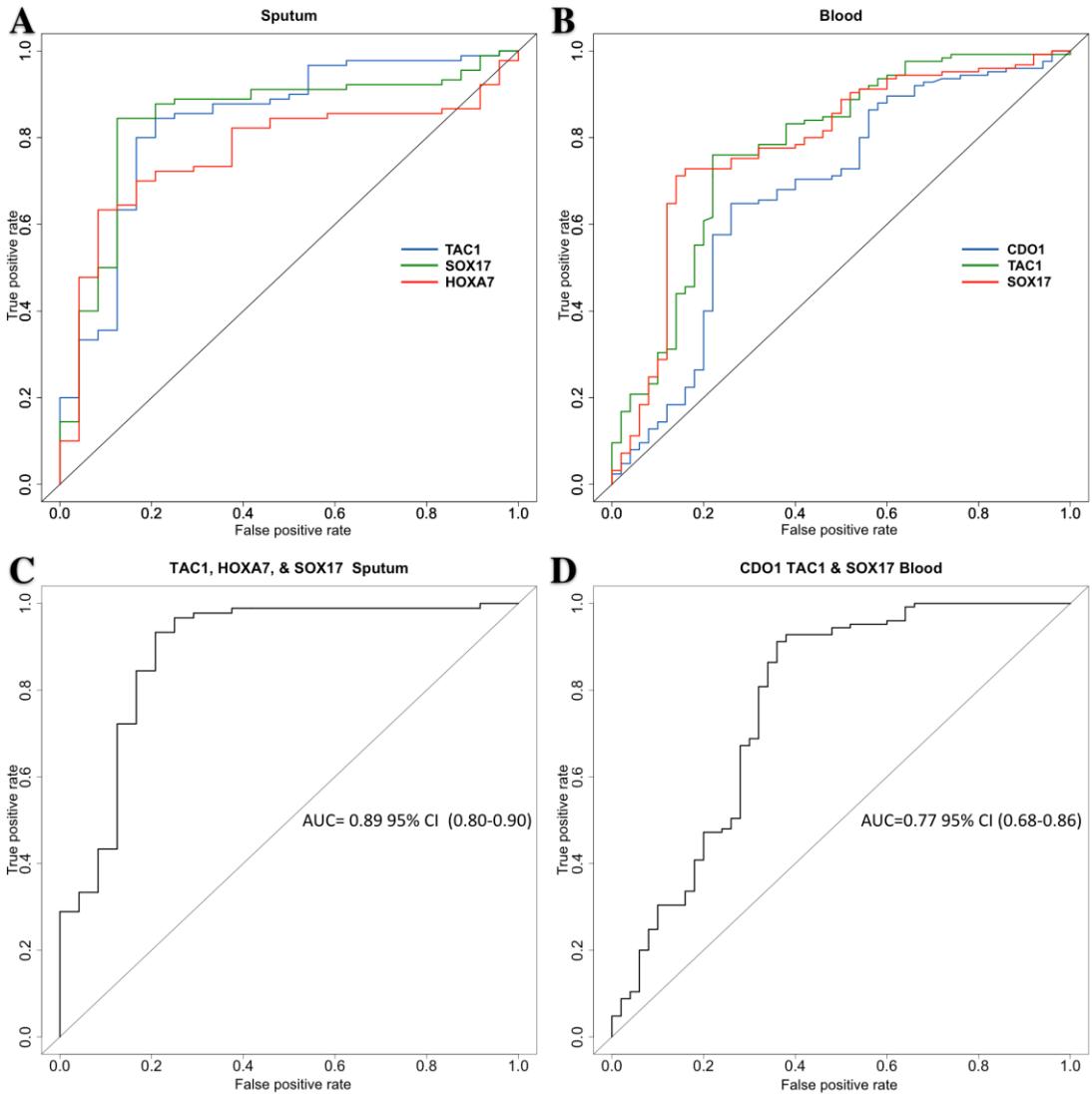

ZPF42
10^{0}
10^{-5}
10^{-10}
10^{-15}
10^{-20}
10^{-25}
Tumor CA
Blood CA
Blood Control
Sputum CA
Sputum Control

Figure 2. Receiver operator classification curves for lung cancer detection.

A. ROC curves comparing the 3 genes with the largest areas under the curve for sputum. **B.** ROC curves comparing the 3 genes with the largest areas under the curve for blood. **C.** ROC of the combined methylation status of the genes from sputum with the largest area under the curve. **D.** ROC of the combined methylation status of the genes from blood with the largest area under the curve.



Abbreviations: area under the curve: AUC, 95 % confidence interval: 95% CI.

Figure 3. Receiver operator classification curves for cancer predictions.

ROC curves assessing the accuracy of the predictions for lung cancer performed on the testing subset by using as predictors the ΔC_t values for all six genes, age, pack-year, COPD status and FVC values. The left plot is obtained using sputum samples, the middle one using blood samples and the right one the ROC curve for the clinical predictors alone.

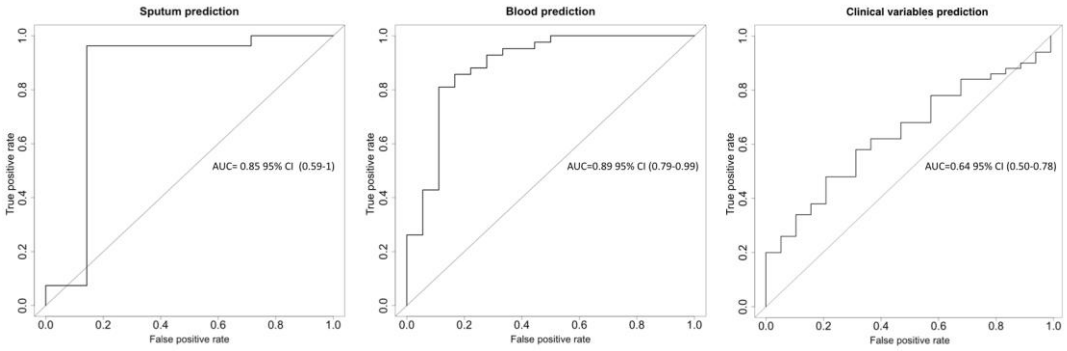
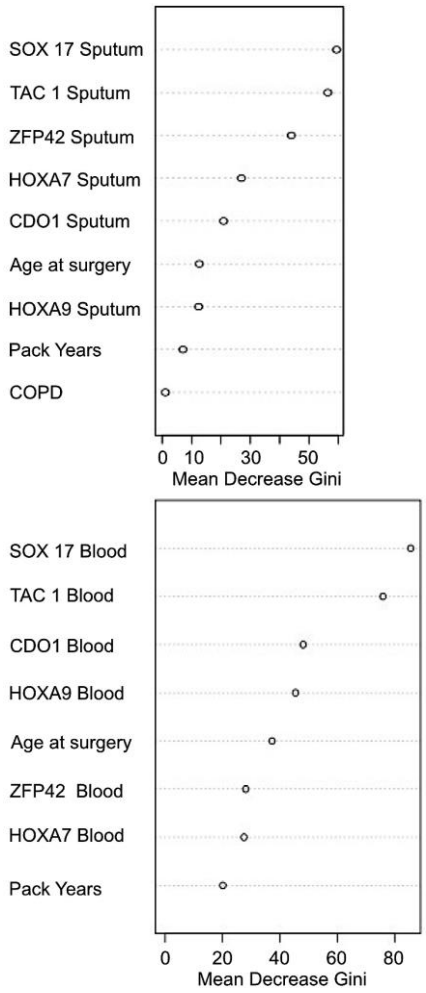


Figure 4. Variable importance plot for random forest prediction[J10]

The plot details the relative importance of each of the variables to the model's accuracy (including: methylation μ $2^{-\Delta CT}$ values, nodule size, age, pack-year, COPD status and FVC values). The x-axis is the mean decrease in the Gini co-efficient that results when that variable is included in the model. The Gini coefficient is a measure of inequality among the trees in the random forest, and in this case represents the performance of the random forest model with and without a variable included. Those variables that have the highest decrease in the Gini coefficient were most likely to create consensus among the individual decision trees used in the model (or reduce inequality) when included in the model. These variables are therefore most predictive of the outcome of the model overall. Those variables with a small decrease in the mean Gini coefficient are relatively less important to the prediction made by the random forest model.



Tables

Table 1. Baseline Characteristics of the 210 Subjects.

Patient Characteristics	Cancer (N=150)	Control (N=60)	*p* Value
Age at surgery (years) (IQR)	68 (62-75)	63 (55-73)	0.007
Gender			
Male (%)	63 (42%)	33 (55%)	0.094
Female (%)	87 (58%)	27 (45%)	
Race			
White (%)	120 (80%)	51 (85%)	
Black (%)	19 (13%)	3 (5%)	0.087
Other (%)	11 (7%)	6 (10%)	
Stage			
IA-IB (%)	136 (91%)	NA	NA
IIA (%)	14 (9%)	NA	
Histology			
Adenocarcinoma (%)	121 (81%)	NA	
Squamous-cell (%)	26 (17%)	NA	NA
Adenosquamous (%)	3 (2%)	NA	
Smoking status			
Current (%)	27 (18%)	7 (12%)	
Former (%)	87 (58%)	34 (57%)	0.176
Never (%)	31 (21%)	19 (32%)	
Pack-year (IQR)	30 (10-50)	20 (0-35)	0.010
COPD (%)	41 (27%)	12 (20%)	0.370
FEV1 % Predicted (IQR)	84 (70-99)	85 (70-100)	0.861
FVC % Predicted (IQR)	92 (80-103)	87 (80-110)	0.682
FEV1/FVC % Ratio (IQR)	73 (68-78)	77 (70-79)	0.080
Nodule size (cm)	2 (1.5-3)	1.5 (1.1-3)	0.01
< 1cm	6 (4%)	13 (22%)	0.001

1-2 cm	52 (35%)	19 (32%)	
> 2 cm	92 (61%)	28 (47%)	
Nodule volume (cm^3)	4.19 (1.77-14-14)	1.6 (0.52-18.12)	0.001

Abbreviations: Chronic obstructive pulmonary disease: COPD, Forced Expiratory Volume in one second: FEV1, Forced vital capacity: FVC, Interquartile range: IQR.
Nodule size % <1cm, 1-2, >2cm

Table 2A Gene Methylation Detection, Sensitivity, Specificity Using Detectable vs. Non-detectable cutoff from Figure 1.

	Cancer (n=90)		Control (n=24)			
Sputum	**n**	**Sensitivity**	**n**	**Specificity**	**PPV**	**NPV**
CDO1	70	78%	8	67%	90%	44%
TAC1	77	86%	6	75%	93%	58%
HOXA7	57	63%	2	92%	97%	40%
HOXA9	84	93%	22	8%	79%	25%
SOX17	76	84%	3	88%	96%	60%
ZFP42	78	87%	9	63%	90%	56%
TAC1, HOXA7, SOX17	88	98%	7	71%	93%	89%

	Cancer (n=125)		Control (n=50)			
Plasma	**n**	**Sensitivity**	**n**	**Specificity**		
CDO1	81	65%	13	74%	86%	46%
TAC1	95	76%	11	78%	90%	57%
HOXA7	42	34%	4	92%	91%	36%
HOXA9	108	86%	27	46%	80%	58%
SOX17	91	73%	8	84%	92%	55%
ZFP42	105	84%	23	54%	82%	57%
CD01, TAC1, SOX17	116	93%	19	62%	86%	78%

Table 2B. Gene Methylation Sensitivity, Specificity, at optimal cutoffs with AUC and Association with Cancer Diagnosis using Sputum and Plasma.

Sputum	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**AUC**	**95% CI**
CDO1	78%	67%	90%	45%	0.70	(0.57 - 0.84)
TAC1	84%	79%	94%	57%	0.84	(0.74 - 0.94)
HOXA7	63%	92%	97%	40%	0.77	(0.67 - 0.86)
HOXA9	77%	42%	83%	32%	0.56	(0.41 - 0.69)
SOX17	84%	88%	96%	59%	0.84	(0.75 -0.94)
ZFP42	88%	62%	90%	58%	0.73	(0.60 - 0.87)
TAC1, HOXA7, SOX17	93%	79%	94%	75%	0.89	(0.80 - 0.98)

Plasma	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**AUC**	**95% CI**
CDO1	65%	74%	86%	46%	0.68	(0.58 - 0.77)
TAC1	76%	78%	90%	57%	0.78	(0.70 - 0.86)

HOXA7	33%	94%	93%	36%	0.60	(0.51 - 0.69)
HOXA9	81%	52%	81%	52%	0.62	(0.52 - 0.73)
SOX17	71%	86%	93%	54%	0.78	(0.70 - 0.86)
ZFP42	81%	58%	83%	55%	0.66	(0.56 - 0.75)
CD01, TAC1, SOX17	91%	64%	86%	74%	0.77	(0.68 - 0.86)

Abbreviations: area under the curve (in the ROC curves): AUC, 95 % confidence interval: 95% CI.

Table 3. Performance for lung cancer diagnosis of the independent blinded random forest prediction models on the testing subset

	Sensitivity	Specificity	PPV	NPV	AUC	95% CI
Prediction from Sputum	93%	86%	96%	75%	0.85	0.59-1
Prediction from Blood	93%	67%	87%	80%	0.89	0.79-0.99
Clinical Predictors alone	84%	26%	75%	38%	0.64	0.50-0.78